CLAIMS

1. A method of detecting duplicates in a set of documents having associated nearest neighbor similarity scores, the method including:

for a particular document in the set of documents, selecting nearest neighbors of the particular document; and

flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are identical.

2. The method of claim 1, further including flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are within a tolerance t of one another.

3. The method of claim 1, wherein the nearest neighbor similarity scores are calculated prior to duplicate detection for a different purpose than the duplicate detection and stored with the documents.

4. The method of claim 1, wherein the k nearest neighbors are determined prior to duplicate detection for a different purpose than the duplicate detection and stored with the documents.

5. The method of claim 1, wherein the documents are text documents.

6. The method of claim 5, wherein the text documents include visual formatting.

7. The method of claim 1, wherein the documents are voice recordings.

8. The method of claim 1, wherein the documents are musical performances.

9. The method of claim 1, wherein the documents are graphic images.

10. The method of claim 1, wherein the nearest neighbors are limited to k nearest neighbors.

11. A method of detecting duplicates in a set of documents, the method including:

identifying nearest neighbors of documents in the set of documents, based on nearest neighbor similarity scores;

for a particular document in the set of documents, flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are identical.

12. The method of claim 11, further including flagging as potential duplicates the nearest neighbors of the particular document that have respective nearest neighbor similarity scores that are within a tolerance t of one another.